Capstone Project - The Battle of Neighborhoods

Introduction/Business Problem

The manager of a chain of niche bakeries has been in touch as the company would like to expand to Scotland, having opened shops across 10 locations across England. They have requested that I compare the postcodes of the Scottish capital, Edinburgh, to inform what location would be best suited for their business. Of particular interest are the existing cafés, bistros and bakeries that would, naturally, compete with their branch for customers, as well as the general layout of the city. The client's key requirement is that the bakery should be centrally located, to maximise foot traffic during busy times such as festivals.

Data

I will use the Foursquare API (https://foursquare.com/) for this analysis:

- The location of interest is Edinburgh.
- Using the latitude and longitude of each postcode in the city, I will explore the city centre, paying specific attention to eateries (cafés, bistros and bakeries). This will involve the use of the explore function and a k-means clustering algorithm.
- Finally, I will use the Folium library to map out the city, showing the clusters of eateries in Edinburgh.

To obtain coordinates for each location, I will use the '2020-2 Scottish Postcode Directory Files' datasets provided by National Records Scotland. This is a list of active and deleted postcodes in Scotland, offered freely online (https://www.nrscotland.gov.uk/statistics-and-data/geography/nrs-postcode-extract).

Methodology

i) Exploratory data analysis

The first step was to download and import the dataset from the provided .csv file. The dataset was found to contain columns describing each of the postcodes in Scotland, their respective districts and sectors, map coordinates among others. However, it was also described that the table included both 'live' postcodes as well as some that are not currently in use. Therefore, the 'DateOfDeletion' column was used to filter out all 'dead' postcodes from the dataset. Also, columns not required for this analysis were all dropped during cleaning, leaving only the name of the postcode, latitude and longitude.

ii) Geo-mapping

Because this analysis involved the use of the Foursquare API to obtain venue information, it was necessary to define a function that would run each postcode against their database to return a list of venues in that specific postcode. In addition to the required criteria such as 'ID' and 'Secret', I introduced 'CategoryId' to the function to limit the venues returned to those of interest, i.e. cafés, bistros and bakeries.

The geopy package was then used to get the exact coordinates of Edinburgh, which were then applied in a second function to narrow the Foursquare search area to a radius of 1000m around that point.

The list of venues returned from the Foursquare calls was then cleaned by dropping duplicates. The remaining venues were then grouped by postcode. One-hot encoding (turning categorical data into binary variables for ease of analysis) was applied, and the venues sorted in order of which type/category is most common in each of the postcodes.

iii) Machine learning algorithm

A K-means clustering algorithm was employed to cluster the list of venues I had obtained by the distinguishing characteristics, using k=5. This number of clusters was selected at random. A cluster label was generated for each row of the data frame, corresponding to the most common venues in that postcode. Their

It was then possible to merge the results data frame (containing the cluster labels) with the earlier columns showing the coordinates of each postcode. This data was used to create a map of Edinburgh using the Folium package, showing the geographical positions of venues, colour-coded by their respective cluster.

Results

On dropping the postcodes not currently in use in Scotland, then limiting the dataset to postcodes within 1 km of Edinburgh, there were 667 postcodes left. The result was 479 venues, many of which were duplicated because of the overlap between postcodes (and, subsequently, Foursquare calls).

Within the results for all relevant postcodes, there were 14 unique categories:

- Bakery
- Café
- Candy Store
- Creperie
- Cupcake Shop
- Dessert Shop
- Gelato Shop
- Gift Shop
- Gourmet Shop
- Ice Cream Shop
- Pastry Shop
- Pie Shop
- Thai Restaurant

Of 667 postcodes included in the search, only 32 had businesses of interest within the search radius. The maximum number of venues in a single postcode was 3, though each of the other 29 postcodes had at least one shop.

As a result of using k=5, there were 5 cluster labels created by the machine learning algorithm:

Table 1. Cluster Labels & Types of Shops in Each

Cluster Label	Key types of Shop/Venue	Colour Legend
0.0	Gourmet Shop	Red
1.0	Ice Cream Shop; Tea Shop	Purple
2.0	Bakery	Blue
3.0	Creperie; Dessert Shop; Tea Room	Green
4.0	Gelato Shop	Orange

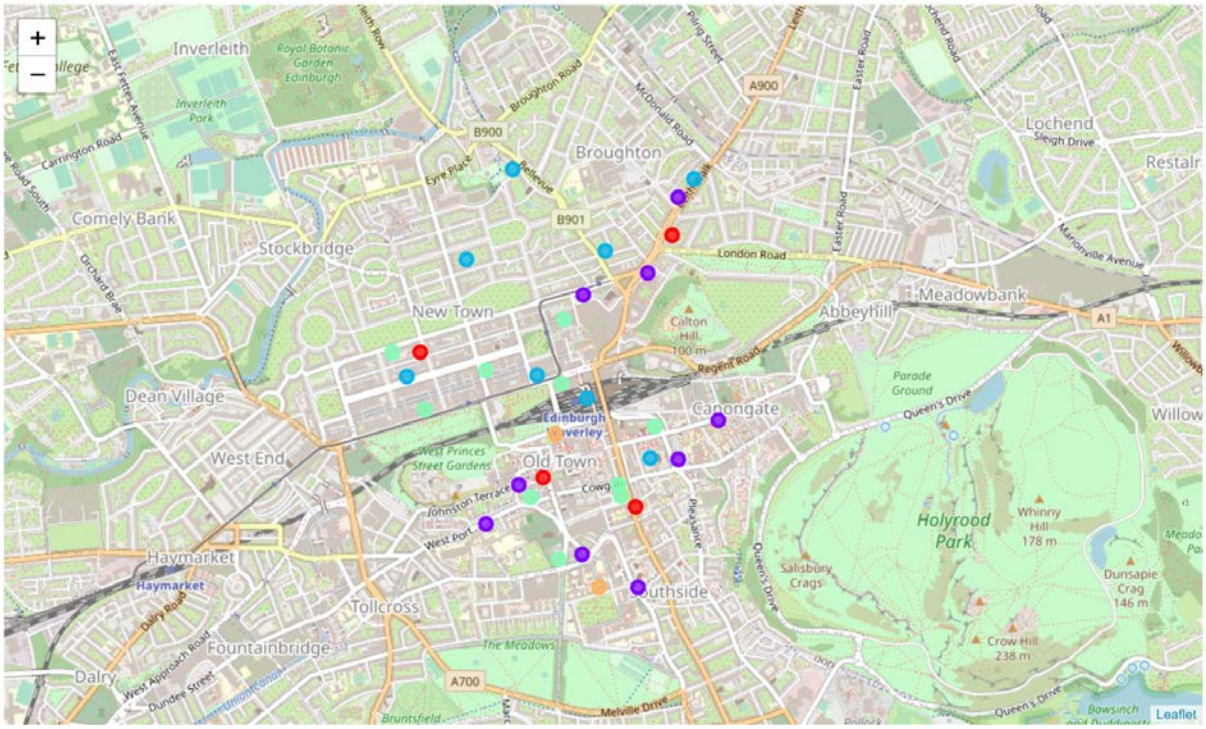

Figure 1 Map of Edinburgh showing the clusters

Discussion

Through the application of multiple Python packages (pandas, NumPy, sci-kit learn, geopy, folium) it was possible to use a dataset of Scottish postcodes to create clusters of baked goods shops in Edinburgh's city centre. Of the five groups, the shops in those labelled '2.0' (blue) and '3.0' (green) are most likely to be direct competitors of my client's business.

The main strengths of this analysis were the use of K-means clustering, an unsupervised machine learning algorithm, for developing the clusters. Also, the dataset used was very detailed, making it easier to obtain accurate venue data from Foursquare.

However, one key challenge that I encountered was in the categorisation of venues in the Foursquare API – of note, 'Thai Restaurant' was one of the categories returned, despite having restricted 'CategoryId' in the original function to shops that would sell the same kind of goods as a bakery. This may have been an oversight by Foursquare, but could also indicate the different functions that specific venues serve, e.g. a Thai restaurant could also sell baked goods!

Another limitation of this analysis was that although in K-means clustering it is acceptable to select 'k' clusters at random, it would have been a more robust method to re-run the algorithm using a variety of numbers as 'k'. It would then be possible to determine which value of 'k' resulted in the least standard error.

Recommendations

The clustered results show that currently, several shops sell baked goods in the Edinburgh City Centre, but for optimal location, the client should open shop away from the blue and green clusters. These shops would present the most significant competition for their customers wishing to buy baked goods. Therefore, an example of a good location for their bakery would be in the area between northern Southside and south-eastern Canongate.

Conclusion

This analysis demonstrates the utility of Python packages and machine learning algorithms in determining a suitable site for a new business. Not only was it possible to make use of the Foursquare API to collect shop locations, but through K-means clustering, I have come up with an analysis to inform my client's business plans.